RESEARCH PAPER

Certainty-based Preference Completion

Lei Li[1,2†], Minghe Xue[2], Zan Zhang[2], Huanhuan Chen[3] & Xindong Wu[1]

[1]Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, Hefei 230009, China

[2]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China

[3]School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China

Keywords: Preference completion; Nondeterministic; Certainty; Subjective probability; Conflict

Citation: Li, L., et al.: Certainty-based preference completion. Data Intelligence 4(1), 112-133 (2022). doi: 10.1162/dint_a_00115
Received: October 15, 2021; Revised: December 7, 2021; Accepted: January 11, 2022

ABSTRACT

As from time to time it is impractical to ask agents to provide linear orders over all alternatives, for these partial rankings it is necessary to conduct preference completion. Specifically, the personalized preference of each agent over all the alternatives can be estimated with partial rankings from neighboring agents over subsets of alternatives. However, since the agents' rankings are nondeterministic, where they may provide rankings with noise, it is necessary and important to conduct the certainty-based preference completion. Hence, in this paper firstly, for alternative pairs with the obtained ranking set, a bijection has been built from the ranking space to the preference space, and the certainty and conflict of alternative pairs have been evaluated with a well-built statistical measurement Probability-Certainty Density Function on subjective probability, respectively. Then, a certainty-based voting algorithm based on certainty and conflict has been taken to conduct the certainty-based preference completion. Moreover, the properties of the proposed certainty and conflict have been studied empirically, and the proposed approach on certainty-based preference completion for partial rankings has been experimentally validated compared to state-of-arts approaches with several datasets.

1. INTRODUCTION

In a preference completion problem, with a set of agents (users) and a set of alternatives (items), each agent (user) has his/her partial ranking over a subset of alternatives (items) and the goal of this problem is to infer each agent (user)'s personalized ranking or preference over all the alternatives (items) including

† Corresponding author: Lei Li (Email: lilei@hfut.edu.cn; ORCID: 0000-0002-5374-7293).

those alternatives (items) the agent (user) has not yet handled. Obviously, from time to time it is impractical to ask agents to provide linear orders over all alternatives, especially in big data environments [1]. For example, perhaps the agent does not know the status of some alternatives because there are too many alternatives, which makes it hard for the agent to rank all of them. Or perhaps some alternatives are incomparable for a certain agent. All these situations mentioned above result in partial rankings, and it is necessary to introduce preference completion.

The preference completion problem has been applied to applications in many areas, such as social choice, and recommender system [2], which can be very useful in community detection [3, 4], or graph anomaly detection [5]. For example, in social choice, each voter (agent) can cast a ballot by a ranking over all candidates (alternatives), or a partial ranking over some candidates (alternatives). As for these partial rankings, it is necessary to form a ranking over all candidates by a certain voting rule. In a recommendation system, each user can rate some items. Then the task of the recommendation system is to predict the rate on the items that have not been rated by him/her. To satisfy this requirement, two common approaches including the matrix factorization approach and the neighborhood-based approach are introduced to handle the preference completion. The traditional algorithms on these two approaches are usually rating-oriented, while a recent line of work focuses on the ranking-oriented algorithms [6, 7] due to the drawbacks of the rating-oriented algorithms. In this paper, we focus on the ranking-oriented neighborhood-based approach.

Traditionally, in neighborhood-based preference completion, it is first to find the near neighbors of each agent and then aggregate these neighbors' rankings to produce the predicted preference by a certain voting rule [6]. However, this task has some inevitable issues. For example, an agent may exhibit irrational behaviors or provide rankings in a noise setting. To address this issue, many rating-oriented trust-based approaches have been proposed with additional contextual information. Meanwhile, the ranking-oriented approach has left much room for better research. Liu et al. [8] proposed an anchor-based algorithm with many other agents' ranking information leveraged to ignore the presence of randomness.

Here in this paper a certainty-based preference completion algorithm is proposed on the basis of Liu's [8] work. More precisely, after finding the k-nearest neighbors by the anchor-kNN algorithm Liu proposed, we use the certainty-based voting algorithm introduced in this paper to complete the preference (ranking) instead of using the traditional majority voting rule. The traditional majority voting rule tends to cause wrong judgment especially when both sides have close votes. In this case, a slight randomness even can cause different outcomes by the majority voting rule. For this reason, this paper introduces a certainty-based voting algorithm to deal with this problem. Importantly, when we take a vote on two alternatives, the certainty which measures the degree that the two alternatives can be preferred or comparable should be introduced. Only when the certainty value satisfies a defined threshold, we can go further to have three-way preference decision instead of assigning 0 or 1 for the two alternatives simply. Hence, the certainty-based voting algorithm avoids the wrong judgment when both sides have close scores or rankings made in a noise setting. In this paper, before formulating the certainty and presenting the certainty-based preference completion algorithm, we consider the certainty and preference space first to introduce the three-way preference between two alternatives.

Technically, in a ranking pool gathered from agents, the rankings including alternative pair A and B can be aggregated to form the preference between A and B. Mathematically, a bijection can be built from the ranking space to the preference space for alternative pair A and B. Here, the ranking space consists of all the partial rankings on A and B from agents, while the preference space consists of three-way preference between A and B, which includes

- preference (prefer A to B, denoted as P^+_{AB}),
- dispreference (prefer B to A, denoted as P^-_{AB}), and
- uncertainty (no preference between A and B, denoted as C^-_{AB}),

according to the trisecting and acting models of human cognitive behaviors [1, 9]. Thus, the following three situations are distinguished:

- The agents prefer alternative A to alternative B, which can be confirmed by high preference P^+_{AB}, low dispreference P^-_{AB}, and low uncertainty C^-_{AB}.
- The agents prefer alternative B to alternative A, which can be confirmed by low P^+_{AB}, high P^-_{AB}, and low C^-_{AB}.
- The agents are uncertain about the preference between alternative pair A and B, i.e., A and B are unpreferred, which can be confirmed by low P^+_{AB}, low P^-_{AB}, and high C^-_{AB}.

It is obvious that when C^-_{AB} is low, the preference between A and B can be determined, i.e., A and B are preferable. Hence, the certainty of preference can be introduced to describe the trustworthiness of the preference, which is denoted as C^+_{AB}, and it can be calculated as $C^+_{AB} = 1 - C^-_{AB}$. The certainty of preference can be taken as the subjective probability of the preference, following the proposition that the certainty is the degree of belief that an individual has on the preference [10]. Hence, in this paper, the certainty can be evaluated based on a well-built statistical measurement, which defines a bijection from ranking space to preference space, enabling the estimation on the pairwise preference with neighbors' partial rankings via mapping them to

$$(\text{preference } P^+_{AB}, \text{dispreference } P^-_{AB}, \text{uncertainty } C^-_{AB}).$$

Our definition on certainty should capture the following key properties:

– **Property 1**: Certainty C^+_{AB} increases as the number of rankings between alternative pair A and B increases for a fixed ratio of rankings from A to B and rankings from B to A.
– **Property 2**: Certainty C^+_{AB} decreases as the extent of conflict increases in the partial rankings between alternative pair A and B.

Our main contributions in this paper can be summarized as follows:

- As pointed out in [11], it is necessary and important to introduce the certainty and conflict of the preference between alternative pairs, and from time to time the certainty and conflict of the preference are more important than the preference itself. In this paper, a probability-based certainty and conflict are introduced under Properties 1 & 2, to describe the trustworthiness of the preference.
- A certainty-based voting algorithm using the certainty and conflict is proposed for conducting the certainty-based preference completion in nondeterministic settings.
- We empirically study the properties of the proposed approach, and experimentally validate the proposed approach compared to the state-of-the-art approaches with several datasets.

This paper is organized as follows. Section 2 reviews existing works on the Plackett-Luce model, Kendall-Tau distance and anchor-*k*NN algorithm. In Section 3, a bijection has been built from ranking space to preference space, and certainty and conflict of alternative pairs have been evaluated based on a well-built statistical measurement. In Section 4, a certainty-based voting algorithm has been taken to conduct the preference completion with the certainty and conflict. In addition, Section 5 studies empirically the properties of the proposed approach about certainty and conflict. Moreover, Section 6 has been experimentally validated compared to the state-of-the-art approaches with several datasets. Finally, Section 7 summarizes this paper and presents the future work.

2. BACKGROUND

2.1 Plackett-Luce Model

Given a set of m alternatives and a set of n agents, let $y(y_1, y_2, \ldots, y_m)$ denotes the latent features of alternatives and $x(x_1, x_2, \ldots, x_n)$ denotes the latent features of agents. Agent i's ranking R_i is determined by a statistical model for ranking data. Hence, as a widely-used statistical model, the Plackett-Luce model [12, 13] is adopted to generate the rankings of agents. In this paper, each alternative is assigned a positive value named utility. The greater this utility is, the more likely its corresponding alternative is ranked at a higher position [14]. In [14], the realized utility for every alternative j on agent i is determined by

$$u_{ij}(x_i, y_j) = \theta(x_i, y_j) + \varepsilon_{i,j}, \tag{1}$$

where $\theta(x_i, y_j)$ is agent i's expected utility on alternative j and can be determined by the closeness of the latent feature x_i and y_j, measured by $\theta(x_i, y_j) = \exp(-||x_i - y_j||_2)$, and $\varepsilon_{i,j}$ is a zero mean independent random variable that follows a Gumbel distribution. When the realized utilities set $u_i(u_{i1}, u_{i2}, \ldots, u_{im})$ of agent i is obtained, agent i ranks the alternatives in a decreasing order according to the realized utilities. After repeating this for n times, synthetic datasets of all the agents can be generated for experiments. For more details, please refer to the following Algorithm 1.

Algorithm 1. Sampling from Plackett-Luce Model.

Input: A latent feature set $x(x_1, x_2, ..., x_n)$ on n agents; a latent feature set $y(y_1, y_2, ..., y_m)$ on m alternatives.
Output: A dataset of rankings $R\{R_1, R_2, \cdots, R_n\}$

for x_i in $x(x_1, x_2, ..., x_n)$ **do**
  for y_j in $y(y_1, y_2, ..., y_m)$ **do**
    Sample $\varepsilon_{i,j}$ follow a Gumbel distribution
    Compute $u_{ij}(x_i, y_j) = \exp(-||x_i - y_j||_2) + \varepsilon_{i,j}$
  end for
end for
for x_i in $x(x_1, x_2, ..., x_n)$ **do**
  for $j = 1$ to m **do**
    Choose an alternative y_j from y with a probability proportional to u_{ij}.
    $R_i \leftarrow R_i > y_j$ and $y \leftarrow y \backslash \{y_j\}$
  end for
end for
return $R\{R_1, R_2, \cdots, R_n\}$

2.2 Kendall-Tau Distance

Given two agents' rankings R_1 and R_2 over the same alternatives, the Kendall-Tau distance can be introduced to measure the similarity of R_1 and R_2, which is the total number of disagreements in pairwise comparisons between alternatives in the linear rankings. For alternative j in R_i, $R_i(j)$ represents the position in R_i. For example, with a ranking of alternatives represented by R_i, if j in R_i is the top-ranked alternative, then $R_i(j) = 1$. The normalized Kendall-Tau distance between R_1 and R_2 is

$$\mathrm{NK}(R_1, R_2) = \frac{\sum_{j_1 \neq j_2 \in R_1} I\left(\prod_{k=1,2}(R_k(j_1) - R_k(j_2)) < 0\right)}{\binom{|R_1|}{2}} \tag{2}$$

where $I(v)$ is an indicator that is set to be 1 if the argument v is true; otherwise, it is set to be 0.

Moreover, if the rankings have not shared completely the same alternatives, the intersection of the two alternative sets can be taken for computing the normalized Kendall-Tau distance.

2.3 Anchor-kNN Algorithm

Before the introduction of the anchor-*k*NN proposed in [8], we first present the idea of KT-*k*NN, which simply uses the Kendall-Tau distance to find the agent's neighbors. If the Kendall-Tau distance between two rankings R_i and R_j is small, the latent feature of the agents x_i and x_j should be close, i.e., the two agents have a similar opinion on alternatives.

As the KT-*k*NN algorithm has not considered that agents' preferences may be nondeterministic or agents' rankings are made in noise setting, different from KT-*k*NN, anchor-*k*NN uses other agents' (named as

anchors) ranking data to determine the closeness of two agents rather than considering the two agents' rankings only. The anchor-*k*NN develops a feature $F_{i,j}$ for agents i and j to represent the Kendall-Tau distance between R_i and R_j, i.e., $F_{i,j} = \mathrm{NK}(R_i, R_j)$. Then for measuring the closeness of two agents denoted as $D_{i,j}$, we use the sum of the difference between $F_{i,t}$ and $F_{j,t}$ to find the k-nearest neighbors, where t is the third agent that belongs to all the other agents except agents i and j.

3. CERTAINTY AND PREFERENCE SPACE

In this section, let us present some preliminary definitions first. For an arbitrary alternatives pair A and B, the certainty can be adopted to describe the trustworthiness of the preference between A and B. Technically, following [15], a Probability-Certainty Density Function (PCDF) can be introduced to capture the subjective probability of the ranking. However, unlike [15], following [16] and [17], in this paper certainty is defined based on the PCDF to satisfy Properties 1 & 2.

3.1 Ranking Space

The ranking space consists of all the weighted partial rankings on the alternative pair A and B from agents, including

- the rankings $\{O_{AB}^{(i)}\}$ where A is ranked ahead of B with weight $w_{AB}^{(i)}$ for the ranking $O_{AB}^{(i)}$, and n_{AB} denotes the accumulated weight of rankings $\{O_{AB}^{(i)}\}$, represented by $n_{AB} = \sum_i w_{AB}^{(i)}$,
- the rankings $\{O_{BA}^{(j)}\}$ where B is ranked ahead of A with weight $w_{BA}^{(j)}$ for the ranking $O_{BA}^{(j)}$, and n_{BA} denotes the accumulated weight of rankings $\{O_{BA}^{(j)}\}$, represented by $n_{BA} = \sum_i w_{BA}^{(j)}$, and
- the unordered ones $\{O_{\overline{AB}}^{(k)}\}$ where A and B are not comparable with weight $w_{\overline{AB}}^{(k)}$ for the ranking $O_{\overline{AB}}^{(k)}$, and $n_{\overline{AB}}$ denotes the accumulated weight of rankings $O_{\overline{AB}}^{(k)}$, represented by $n_{\overline{AB}} = \sum_k w_{\overline{AB}}^{(k)}$. Obviously, we have $w_{\overline{AB}}^{(k)} = w_{\overline{BA}}^{(k)}$, and $O_{\overline{AB}}^{(k)} = O_{\overline{BA}}^{(k)}$.

Moreover, the weight $w_{AB}^{(i)}$ for $O_{AB}^{(i)}$ means the quality of ranking $O_{AB}^{(i)}$. Without additional knowledge, we assign $w_{AB}^{(i)}$ to be 1.

Definition 1. Ranking space

$$O = \{< n_{AB}, n_{BA}, n_{\overline{AB}} > | \min\{n_{AB}, n_{BA}, n_{\overline{AB}}\} > 0\}.$$

3.2 Preference Space

Traditionally, the uncertainty is usually ignored, and sometimes dispreference has not been taken into account as well, which leads to some disturbing results shown in empirical study section. According to the trisecting and acting models of human cognitive behaviors [9, 18], the preference space consists of three-way preference between alternatives, which includes

- preference P_{AB}^{+} (prefer A to B),
- dispreference P_{AB}^{-} (prefer B to A), and
- uncertainty C_{AB}^{-} (no preference between A and B).

Definition 2. Preference space

$$P = \{< P_{AB}^{+}, P_{AB}^{-}, C_{AB}^{-} > | \, P_{AB}^{+} + P_{AB}^{-} + C_{AB}^{-} = 1, \min\{P_{AB}^{+}, P_{AB}^{-}, C_{AB}^{-}\} > 0\}.$$

3.3 Certainty of Rankings in Alternative Pairs

The Bayesian inference [19, 20] here is adopted to update the probability with the available contextual information about the rankings in alternative pairs, i.e., update the prior distribution to the posterior distribution [21, 22]. Currently, the offline Bayesian inference has been utilized in this paper. The Bayesian inference can also be applied to online/streaming scenario [23, 24].

Let x_{AB}, x_{BA} and $x_{\overline{AB}}$ be the probability of rankings $\{O_{AB}^{(i)}\}$, $\{O_{BA}^{(j)}\}$ and $\{O_{\overline{AB}}^{(k)}\}$, respectively, where $x_{\overline{AB}} = 1 - x_{AB} - x_{BA}$ and $X = <x_{AB}, x_{BA}>$. In addition, $x_{AB} \in [0,1]$, $x_{BA} \in [0,1]$ and $x_{\overline{AB}} \geq 0$, and thus we then have $x_{AB} + x_{BA} \leq 1$.

Without any additional information, the prior distribution $f(X|O)$ is a uniform distribution. As the cumulative probability of a distribution within [0,1] equals 1, the density of a PCDF has the mean value 1 within [0,1], and this makes $f(X|O) = 1$.

As the ranking sample O conforms to a multinomial distribution [16, 22], we have

$$f(O) = \frac{6(x_{AB})^{n_{AB}}(x_{BA})^{n_{BA}}(x_{\overline{AB}})^{n_{\overline{AB}}}}{n_{AB}!n_{BA}!(n_{\overline{AB}})!} \tag{3}$$

As for posterior distribution $f(O|X)$, it can be estimated as [16, 22]:

$$f(O \mid X) = \frac{f(X \mid O)f(O)}{\int_0^1 f(X \mid O)f(O)\mathrm{d}X} = \frac{(x_{AB})^{n_{AB}}(x_{BA})^{n_{BA}}(x_{\overline{AB}})^{n_{\overline{AB}}}}{\int_0^1 (x_{AB})^{n_{AB}}(x_{BA})^{n_{BA}}(x_{\overline{AB}})^{n_{\overline{AB}}}\mathrm{d}X} \tag{4}$$

Then, the certainty can be determined by the deviations of posterior distribution from the prior distribution, i.e., uniform distribution. Hence, we have the following definition about certainty.

Definition 3. The certainty C_{AB}^{+} of rankings $\{< n_{AB}, n_{BA}, n_{\overline{AB}} >\}$ can be estimated as

$$C_{AB}^{+} = \frac{1}{2}\int_0^1 | \, f(O \mid X) - f(X \mid O) \, | \, \mathrm{d}X = \frac{1}{2}\int_0^1 \left| \frac{(x_{AB})^{n_{AB}}(x_{BA})^{n_{BA}}(x_{\overline{AB}})^{n_{\overline{AB}}}}{\int_0^1 (x_{AB})^{n_{AB}}(x_{BA})^{n_{BA}}(x_{\overline{AB}})^{n_{\overline{AB}}}\mathrm{d}X} - 1 \right| \mathrm{d}X \tag{5}$$

where $\frac{1}{2}$ is to remove the double counting of the deviations.

From this definition, we have $C_{AB}^{+} = C_{BA}^{+}$.

3.4 Conflict of Rankings in Alternative Pairs

The conflict can be determined by the relative difference between weighted rankings n_{AB} and n_{BA}, as in [17]. More specifically,

- there is the largest conflict, when weighted rankings $n_{AB} = n_{BA}$;
- there is the smallest conflict, when weighted rankings $n_{AB} = 0$ or $n_{BA} = 0$.

Hence, we have the following definition about conflict.

DEFINITION 4. The conflict c_{AB} of rankings $\{< n_{AB}, n_{BA}, n_{\overline{AB}} >\}$ can be estimated as

$$c_{AB} = \min\left\{\frac{n_{AB}}{n_{AB}+n_{BA}}, \frac{n_{BA}}{n_{AB}+n_{BA}}\right\} \tag{6}$$

From this definition, we have $c_{AB} = c_{BA}$.

3.5 Bijection from Ranking Space to Preference Space

With Definitions 1, 2, 3 and 4, the following definition can be introduced.

DEFINITION 5. The bijection from ranking space $\{< n_{AB}, n_{BA}, n_{\overline{AB}} >\}$ to preference space $\{< P^{+}_{AB}, P^{-}_{AB}, C^{-}_{AB} >\}$ can be estimated as

$$P^{+}_{AB} = \frac{n_{AB}}{n_{AB}+n_{BA}+n_{\overline{AB}}} C^{+}_{AB} \tag{7}$$

$$P^{-}_{AB} = \frac{n_{BA}}{n_{AB}+n_{BA}+n_{\overline{AB}}} C^{+}_{AB} \tag{8}$$

$$C^{-}_{AB} = 1 - C^{+}_{AB} \tag{9}$$

4. CERTAINTY-BASED PREFERENCE COMPLETION

This section proposes the certainty-based preference completion approach. The framework of our approach is shown in Figure 1. It includes two processes. One is to find the *k*-nearest neighbors for user *i* with the anchor-*k*NN algorithm Liu [8] proposed. The other one is to conduct a linear ranking for user *i* over all alternatives. In this section, we focus on the latter one. As for the latter one, with the neighbors' partial ranking, a certainty-based voting algorithm is introduced to estimate pairwise preference for all pair alternatives, and then these pairwise preferences can form a linear ranking for the user *i*.

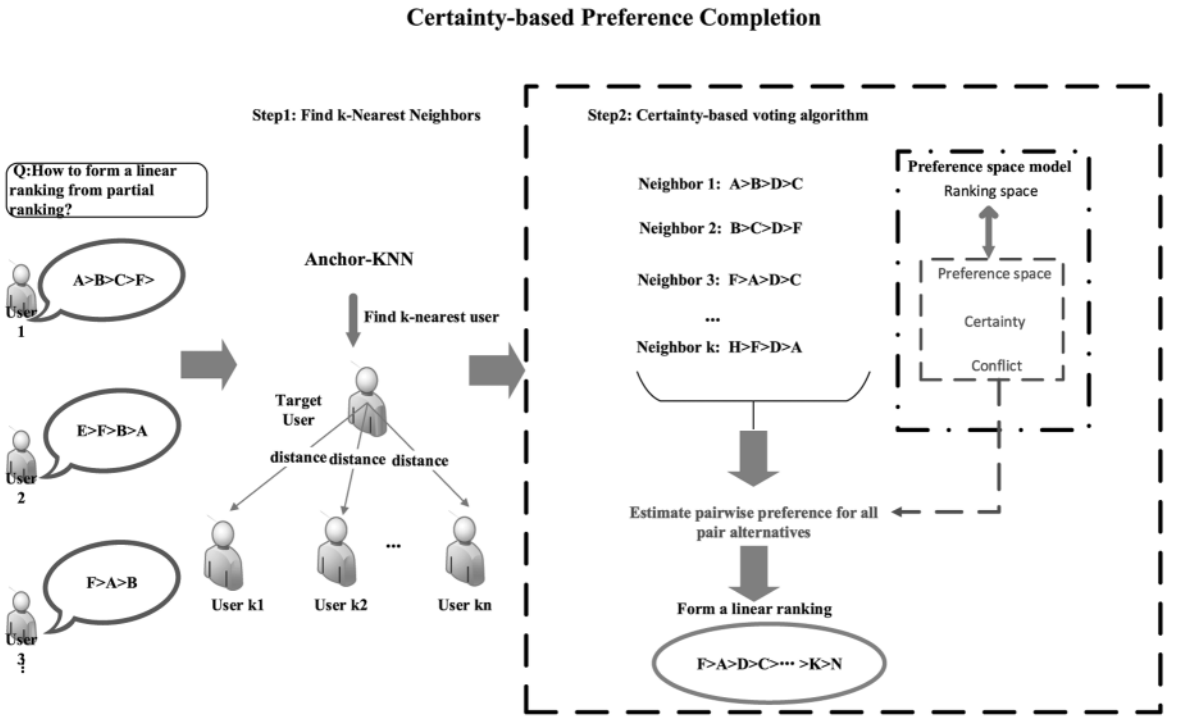

Figure 1. Certainty-based preference completion process.

4.1 Certainty-based Voting Algorithm

First, let us introduce a definition.

DEFINITION 6. With preference space $\{< P_{AB}^{+}, P_{AB}^{-}, C_{AB}^{-} >\}$, the following conclusions can be obtained:

- if uncertainty $C_{AB}^{-} \geq \varepsilon_1$, alternatives A and B are unpreferred;
- if $C_{AB}^{-} < \varepsilon_1$,
  - if $P_{AB}^{+} - P_{AB}^{-} \geq \varepsilon_2$, user i prefers A to B;
  - if $P_{AB}^{-} - P_{AB}^{+} \geq \varepsilon_2$, user i prefers B to A;
  - otherwise, A and B are unpreferred;

where ε_1 and ε_2 are thresholds to rule out the fuzziness of comparison.

In the existing work, with the rankings of neighbors obtained by k-nearest neighbors algorithm, common voting rules[①], such as majority voting, can be taken to estimate pairwise preference for conducting the preference completion.

[①] common voting rules may include positional scoring rules, maximin, and Bucklin. For more details, please refer to [21].

In contrast, in this paper, we use a certainty-based voting rule with certainty and conflict to obtain pairwise preference. The certainty and conflict measure the trustworthiness that the pair alternatives can be preferred or comparable. If the certainty satisfies a defined threshold, we can then evaluate the degree that the user i prefers one to another denoted by P^+_{AB} and P^-_{AB}. Then, only if the difference between the two-way preference has reached a value, we can make a preference decision on the two alternatives. Technically, for the alternative pair A and B with $C^-_{AB} < \varepsilon_1$ and $\left|P^+_{AB} - P^-_{AB}\right| \geq \varepsilon_2$, a preference decision between A and B can be made. The process for estimating pairwise preference is also shown in Algorithm 2. We apply this algorithm on all alternative pairs, and then we get all the pairwise preferences.

Algorithm 2. Certainty-based voting algorithm for estimating pairwise preference.

Input: A pair of alternatives A and B, neighbors rankings $R_1, R_2, ..., R_k$
Output: pairwise preference for alternative A and B denoted by ψ_{AB}.

for R_i in $\{R_1, R_2, ..., R_k\}$ **do**
  if A in R_i and B in R_i **then**
    $n_{AB} \leftarrow n_{AB} + 1$ or $n_{BA} \leftarrow n_{BA} + 1$, if A is ranked before or after B
  else
    $n_{\overline{AB}} \leftarrow n_{\overline{AB}} + 1$,
  end if
end for
Compute C^-_{AB} and P^+_{AB}, P^-_{AB} using n_{AB}, n_{BA} and $n_{\overline{AB}}$ with Equation (5), Equation (7), Equation (8) and Equation (9).
if **then**

else

end if
if $C^-_{AB} \geq \epsilon_1$ **then**
  $\psi_{AB} = 0$
end if
if $C^-_{AB} < \epsilon_1$ **then**
  if $P^+_{AB} - P^-_{AB} \geq \epsilon_2$ **then**
    $\psi_{AB} = 1$
  else if $P^-_{AB} - P^+_{AB} \geq \epsilon_2$ **then**
    $\psi_{AB} = -1$
  else
    $\psi_{AB} = 0$
  end if
end if
return ψ_{AB}

4.2 Greedy Order Algorithm

Next, let us combine all the pairwise preferences to form a linear ranking over all alternatives. One possible approach is the greedy order algorithm [25]. This algorithm follows a greedy idea: the order algorithm always picks the alternative that currently has the maximum potential value in the alternatives pool I and ranks it above all the other remaining items. Here, for item i, the potential value v_i is equal to $\sum_{j\in I}\psi_{i,j} - \sum_{j\in I}\psi_{j,i}$. This value aggregates all the pairwise preferences obtained in the previous subsection and represents the preference for item i among all the neighbors' rankings. Then it deletes the picked one from the alternatives pool and updates the potential values of the remaining items by removing the effects of the picked one. Repeat the picking process until the alternatives pool is empty, and then a linear ranking for user i is produced. See Algorithm 3.

Algorithm 3. Greedy order algorithm.

Input: An alternative set Y, neighbors rankings $R_1, R_2, ..., R_k$, all pairwise preferences set $\Psi(\psi_{1,2}, \psi_{1,3}, ..., \psi_{m-1,m})$
Output: A complete ranking R for target user i

```
/*compute the potential value for every alternative*/
for all i ∈ Y do
    v_i = Σ_{j∈I} ψ_{i,j} − Σ_{j∈I} ψ_{j,i},
end for
while I is not empty do
    t = argmax_{t∈I} v_t
    R̂(t) = |I|
    I = I − t
    for all i ∈ Y do
        v_i = v_i + ψ_{t,i} − ψ_{i,t}
    end for
end while
return R
```

5. EMPIRICAL STUDIES ON PROPERTIES OF CERTAINTY

In this section, we study the properties of certainty and conflict in our proposed model.

5.1 Increasing Rankings with Fixed Conflict

Figure 2 plots how certainty C^{+}_{AB} varies with weighted rankings n_{AB} and $n_{\overline{AB}}$ under fixed conflict c_{AB}.

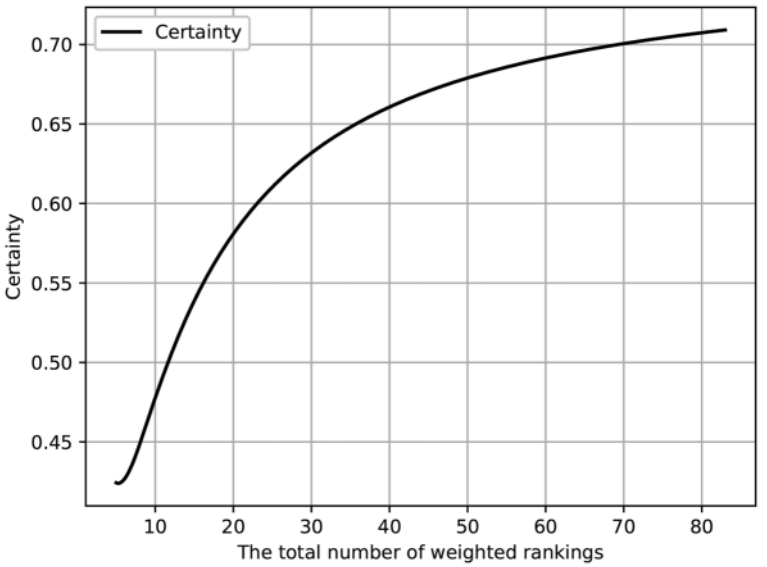

Figure 2. Certainty increases with $n_{AB} + n_{BA}$ when $\frac{n_{AB}}{n_{AB}+n_{BA}}$ and $n_{\overline{AB}}$ is fixed.

This should confirm Property 1.

THEOREM 1. As for fixed $\frac{n_{AB}}{n_{AB}+n_{BA}}$ and $n_{\overline{AB}}$, the certainty C^{+}_{AB} increases with $n_{AB} + n_{BA}$.

Proof: Let $\frac{n_{AB}}{n_{AB}+n_{BA}} = \alpha$, $n_{AB} + n_{BA} = \beta$, and

$$f(\bullet) = \frac{(x_{AB})^{n_{AB}}(x_{BA})^{n_{BA}}(1-x_{AB}-x_{BA})^{n_{\overline{AB}}}}{\int_0^1 (x_{AB})^{n_{AB}}(x_{BA})^{n_{BA}}(1-x_{AB}-x_{BA})^{n_{\overline{AB}}}\,\mathrm{d}X} \tag{10}$$

Then we have

$$C^{+}_{AB} = \frac{1}{2}\int_0^1 |f(\bullet)-1|\,\mathrm{d}X \tag{11}$$

As in [17], x_1, x_2, x_3, x_4 can be defined, such that $f(x_1) = f(x_2) = f(x_3) = f(x_4) = 1$ and

$$C^{+}_{AB} = \int_{x_1}^{x_2}\int_{x_3}^{x_4}[f(\bullet)-1]\mathrm{d}x_{AB}\mathrm{d}x_{BA} \tag{12}$$

where x_1, x_2, x_3, and x_4 are functions of β. Then

$$\begin{aligned}\frac{\partial C^{+}_{AB}}{\partial \beta} &= \frac{\partial x_2}{\partial \beta}\int_{x_3}^{x_4}[f(x_2)-1]\mathrm{d}x_{AB} - \frac{\partial x_1}{\partial \beta}\int_{x_3}^{x_4}[f(x_1)-1]\mathrm{d}x_{AB} + \int_{x_1}^{x_2}\frac{\partial}{\partial \beta}\int_{x_3}^{x_4}[f(\bullet)-1]\mathrm{d}x_{AB}\mathrm{d}x_{BA} \\ &= \int_{x_1}^{x_2}\frac{\partial}{\partial \beta}\int_{x_3}^{x_4}[f(\bullet)-1]\mathrm{d}x_{AB}\mathrm{d}x_{BA}\end{aligned} \tag{13}$$

where

$$\frac{\partial}{\partial\beta}\int_{x_3}^{x_4}[f(\bullet)-1]\mathrm{d}x_{AB}=\frac{\partial x_4}{\partial\beta}[f(x_4)-1]-\frac{\partial x_3}{\partial\beta}[f(x_3)-1]+\int_{x_3}^{x_4}\frac{\partial}{\partial\beta}[f(\bullet)-1]\mathrm{d}x_{AB}$$
$$=\int_{x_3}^{x_4}\frac{\partial}{\partial\beta}[f(\bullet)-1]\mathrm{d}x_{AB} \tag{14}$$

Following Lemma 9 in [17], we have

$$\frac{\partial}{\partial\beta}\int_{x_3}^{x_4}[f(\bullet)-1]\mathrm{d}x_{AB}>0 \tag{15}$$

With Equation (13), we have

$$\frac{\partial C_{AB}^{+}}{\partial\beta}>0 \tag{16}$$

This confirms the results of Theorem 1.

5.2 Increasing Conflict with Fixed Rankings

Figure 3 plots how certainty C_{AB}^{+} varies with weighted rankings n_{AB} and $n_{\overline{AB}}$ under the fixed summation of n_{AB} + n_{BA} and the fixed $n_{\overline{AB}}$. This should confirm Property 2.

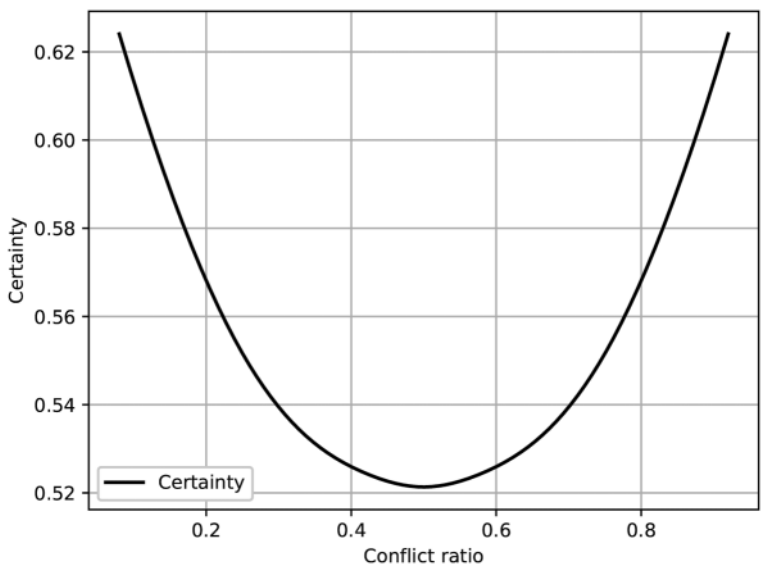

Figure 3. Certainty is concave when $n_{AB}+n_{BA}+n_{\overline{AB}}$ and $n_{\overline{AB}}$ is fixed, and the minimum occurs at $n_{AB}=n_{BA}$.

THEOREM 2. As for fixed $n_{\overline{AB}}$, the certainty C^{+}_{AB} is decreasing with $n_{AB} \leq n_{BA}$, and increasing with $n_{AB} \geq n_{BA}$.

Proof: The details of validation process can be omitted here, as it is similar to one in the proof of Theorem 1. More specifically, with removing the absolute sign and then differentiating it, it can be proved that the derivation is negative for $n_{AB} \leq n_{BA}$, and positive for $n_{AB} \geq n_{BA}$.

6. EXPERIMENTS

In this section, we examine the empirical performance of the certainty-based preference completion algorithm. In the experiments, we compare our certainty-based preference completion algorithm with the common majority voting algorithm [8] and the classic collaborative filtering algorithm (CF) [26]. Both our certainty-based preference completion algorithm and majority voting algorithm use the anchor-*k*NN algorithm to find *k*-nearest neighbors' rankings and utilize these rankings to conduct the preference completion of the target user. While the collaborative filtering algorithm is a rating-oriented algorithm different from the other two. It computes user's similarity to find user's neighbors, and uses their ratings to generate item prediction.

6.1 Datasets

The experiments adopt two forms of datasets to evaluate algorithms' performance.

- One type of dataset is the synthetic one created by the sampler using a Plackett-Luce model with Algorithm 1. The produced synthetic dataset has over 20,000 rankings from agents on the set of 20 alternatives. Each ranking follows a Gumbel distribution.
- The other type of dataset is the Flixster dataset that collects the movie ratings by users with social trust. It has over 8,000,000 ratings on over 2,000 movies. For the experiments, we convert the ratings to rankings, and select over 9,000 rankings on over 50 movies.

6.2 Evaluation Metrics

We evaluate the performance on three metrics: (a) Prediction error, (b) Spearman correlation coefficient, (c) Kendall rank correlation coefficient. The first one measures the quality of the predicted ranking, and the others measure the degree of correlation on the predicted ranking with the original one. Please refer to Pearson [27] and Liu et. al. [2] for more details.

- **Evaluation Metric 1**: This evaluation metric estimates the accuracy on the predicted ranking with the original true one.

$$\Phi_{Prdiction\ Error} = \frac{1}{M} \sum_{X_{i,j,k}=1} I^{-}(Y_{i,j,k}) \tag{17}$$

where M is the maximum of the pairwise error, $Y_{i,j,k} = 1$ means in predicted ranking, alternative user i prefers alternative j to alternative k and $X_{i,j,k} = 1$ represents alternative user i prefers alternative j to alternative k in original ranking. $I^-(v)$ equals 1 when $v < 0$, and equals 0 otherwise.

- **Evaluation Metric 2**: The Spearman correlation coefficient measures the difference of the position for every alternative in predicted ranking and the original one to evaluate the similarity between the predicted ranking and the original one. The greater its value, the more precise our predicted ranking.

$$\Phi_{Spearman\ CC} = \frac{\sum_i (x_i - \overline{x})(y_i - y)}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} \tag{18}$$

to simplify, we have

$$\Phi_{Spearman\ CC} = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \tag{19}$$

where d_i represents the difference on the position of alternative i with the predicted ranking and the original one.

- **Evaluation Metric 3**: The Kendall rank correlation coefficient is very similar to the above evaluation Metric 2, except that it uses the Kendall distance to measure the correlation:

$$\Phi_{Kendall\ CC} = 1 - \frac{4\sum I^-(X_{i,j,k} Y_{i,j,k})}{\left|I_x \cap I_y\right| \cdot \left|(I_x \cup I_y) - 1\right|} \tag{20}$$

where the symbol in Equation (20) has the same meaning with the evaluation Metric 1, I_x represents the alternatives set in original ranking, and I_y represents the alternatives set in predicted ranking.

6.3 Experimental Results on Synthetic Dataset and Flixster Dataset

In this section, we conduct the experiments on a synthetic dataset and the Flixster dataset. With the evaluation metrics separately, the comparison results with different approaches can be presented. The prediction error measures the difference in pairwise preference with the predicted ranking and original ranking. The goal is to reduce the prediction error as far as possible. While the Spearman correlation coefficient and the Kendall rank correlation coefficient measure the similarity between the predicted ranking and the original ranking. We expect the values on these two evaluation metrics can be higher possibly.

(a) **Synthetic dataset**

- As shown in Figure 4, it is very clear that the prediction error tends to be smaller when using certainty-based algorithm than the CF algorithm and the majority voting algorithm. In addition, the two ranking-oriented approaches outperform the rating-oriented approach. For one thing, the ranking contains more preference relation information over alternatives than rating score, and thus it may be easier and more accurate in finding the user's neighbors and completing preference. As a result, the ranking-oriented approach has a lower prediction error. For another, the comparison between the

certainty-based voting algorithm and the majority voting algorithm shows the superiority of the certainty-based one. The preference completion algorithm with certainty considered does reduce the effect of randomness.

- Figure 5(a) shows the performance of Spearman correlation coefficient. On this evaluation metric, the certainty-based voting algorithm performs better than the other two algorithms. This is because our approach with preference space and certainty considered can filter out those pair preferences which have close votes and have lower certainty. This behavior causes the predicted rank much more trustworthy.
- Figure 5(b) shows the performance of Kendall rank correlation coefficient. We can get a similar conclusion with the Spearman correlation coefficient in Figure 5(a), so we do not repeat explanation here.

Roughly speaking, from the experiments on the synthetic dataset, we verify the effectiveness of our proposed certainty-based preference completion algorithm.

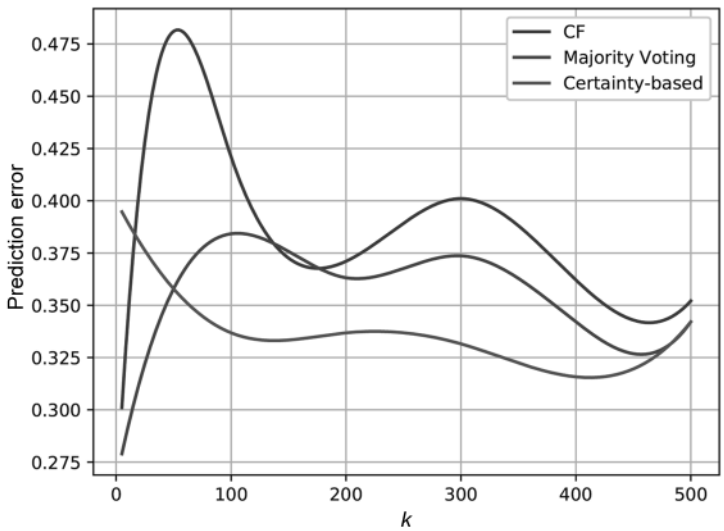

Figure 4. Prediction error on synthetic dataset: *x*-axis denotes the number of neighbors. Plots show the prediction error. For this evaluation metric, smaller values are better.

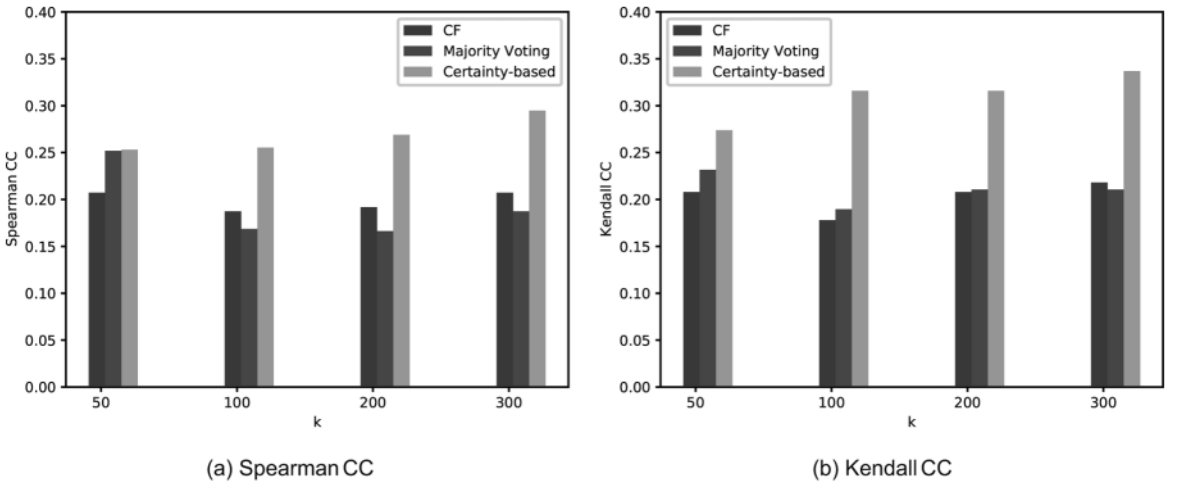

Figure 5. Performance on synthetic dataset: *x*-axis denotes the number of neighbors. Plots show the Spearman correlation coefficient (Spearman CC) and Kendall rank correlation coefficient (Kendall CC). For both evaluation metrics, higher values are better.

(b) **Flixster dataset** The performance of the three approaches is examined on a real-world dataset, Flixster dataset, which contains the rating information. Because the proposed algorithm and the majority voting algorithm both use the anchor-kNN algorithm which need ranking data instead of rating data, we need to convert rating data to ranking data first.

- As shown in Figure 6, when the number of neighbors, $k > 300$, our approach outperforms the other two and the ranking-oriented method still performs better than the rating-oriented method. While when $k < 300$, the result does not perform as expected. A possible reason may be that the process of converting ranking data to rating data inevitably brings errors on the pairwise preference. With more neighbors considered, our proposed algorithm shows its superiority. Thus, the prediction error descends when the number of neighbors grows.
- In Figure 7(a), as we can observe, the certainty-based approach outperforms the other two approaches significantly. This shows a consistent result with the experiments on the synthetic dataset.
- Figure 7(b) shows the a similar performance with Figure 7(a).

In general, with the experiments on the synthetic dataset and Flixster dataset, we can come to a conclusion that the experiments validate our proposed certainty-based preference completion algorithm.

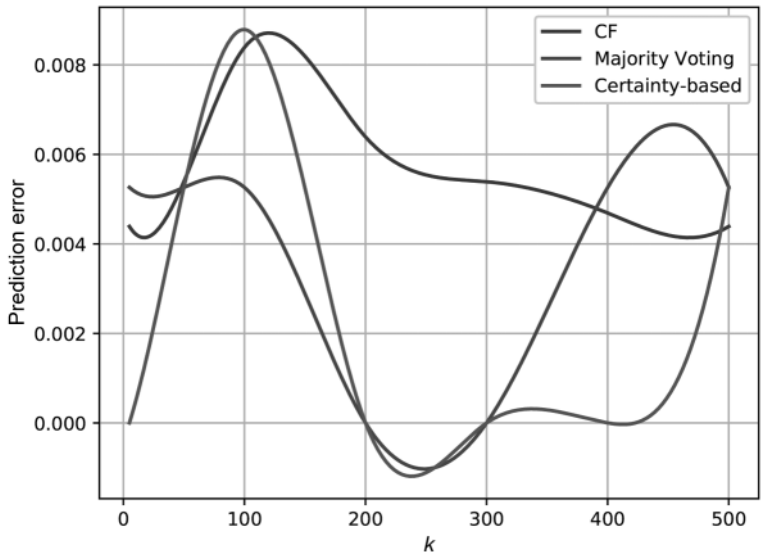

Figure 6. Prediction error on Flixster dataset: *x*-axis denotes the number of neighbors. Plots show the prediction error. For this evaluation metric, smaller values are better.

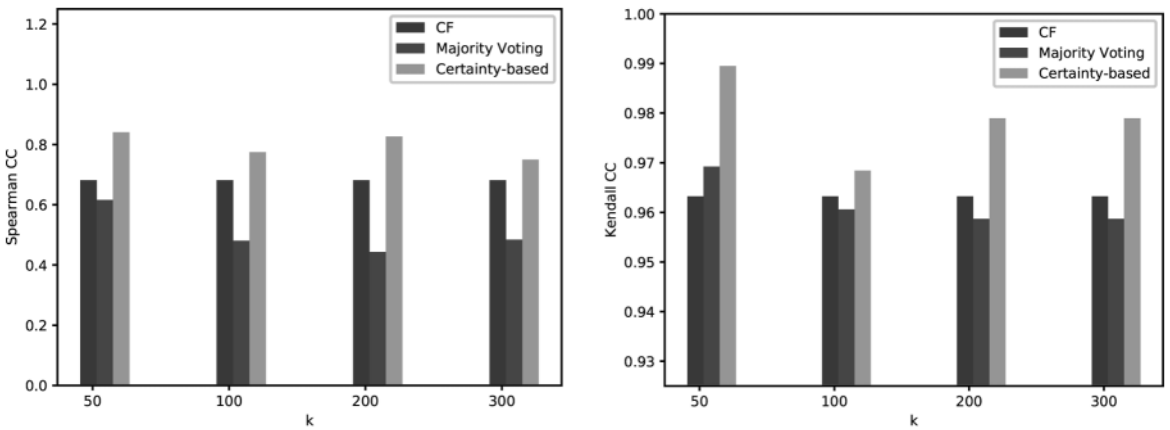

Figure 7. Performance on Flixster dataset: *x*-axis denotes the number of neighbors. Plots show the Spearman correlation coefficient (Spearman CC) and Kendall rank correlation coefficient (Kendall CC). For both evaluation metrics, higher values are better.

7. CONCLUSION AND FUTURE WORK

Due to the fact that the agents' rankings are nondeterministic, where they may provide their rankings under noisy environments, it is necessary and important to conduct the certainty-based preference completion. Hence, in this paper firstly, for alternative pairs a bijection has been built from the ranking space to the preference space, and its certainty and conflict have been evaluated based on a well-built statistical measurement Probability-Certainty Density Function. Then, a certainty-based voting algorithm based on the certainty and conflict has been taken to conduct the preference completion. More specifically, the ranking with high certainty and low conflict can be obtained with the proposed algorithm to conduct the preference completion. Moreover, the properties of the proposed approach about certainty and conflict have been studied empirically, and the proposed approach has been experimentally validated compared to the state-of-the-art approaches with several datasets.

As in real applications, the data is usually unbalanced [28], i.e., some alternative pairs have a lot of rankings, while others only have a few rankings. In our future work, we will propose algorithms to handle unbalanced preference completion both effectively and efficiently.

ACKNOWLEDGEMENTS

This work has been supported by the National Natural Science Foundation of China (No. 62076087, No. 61906059 & No. 62120106008) and the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education of China under grant IRT17R32.

The first author would like to thank his wife Jun Zhang, his parents and friends during his fight with lung adenocarcinoma. "I leave no trace of wings in the air, but I am glad I have had my flight."

AUTHOR CONTRIBUTIONS

All authors including L. Li (lilei@hfut.edu.cn), M.H. Xue (18856337539@163.com), Z. Zhang (zanzhang@hfut.edu.cn), H.H. Chen (hchen@ustc.edu.cn), and X.D. Wu (xwu@hfut.edu.cn) took part in writing the paper. In addition, L. Li designed the algorithm and experiments, and provided the funding; M.H. Xue designed and conducted experiments, and analyzed the data; Z. Zhang analyzed the data.

REFERENCES

[1] Li, L., et al.: Weighted partial order oriented three-way decisions under score-based common voting rules. International Journal of Approximate Reasoning 123(2020), 41–54 (2020)

[2] Liu, T.: Learning to rank for information retrieval. Berlin, Springer (2011)

[3] Liu, F., et al.: Deep learning for community detection: Progress, challenges and opportunities. In: Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI 2020), pp. 4981–4987 (2020)

[4] Su, X., et al.: A comprehensive survey on community detection with deep learning. arXiv preprint arXiv:2105.12584 (2021)

[5] Ma, X., et al.: A comprehensive survey on graph anomaly detection with deep learning. arXiv preprint arXiv:2106.07178 (2021)

[6] Katz-Samuels, J., Scott, C.: Nonparametric preference completion. In: Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS 2018), pp. 632–641 (2018)

[7] Liu, N., Yang, Q.: Eigenrank: A ranking-oriented approach to collaborative filtering. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 83–90 (2008)

[8] Liu, A., et al.: Near-neighbor methods in random preference completion. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2019), pp. 4336–4343 (2019)

[9] Yao, J.: Three-way granular computing, rough sets, and formal concept analysis. International Journal of Approximate Reasoning 116, 106–125 (2020)

[10] Li, L., Wang, Y.: Context based trust normalization in service-oriented environments. In: Proceedings of the IEEE Conference on Autonomic and Trusted Computing, pp. 122–138 (2010)

[11] Hallinan, J.T.: Why we make mistakes. Broadway Books, Portland (2010)

[12] Luce, R.: Individual choice behavior: A theoretical analysis. Dover Publications, New York (1959)

[13] Plackett, R.: The analysis of permutations. Applied Statistics 24, 193–202 (1975)

[14] Liu, A., et al.: Learning plackett-luce mixtures from partial preferences. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2019), pp. 4328–4335 (2019)

[15] Jøsang, A.: A subjective metric of authentication. In: Proceedings of the 5th European Symposium on Research in Computer Security (ESORICS 98), pp. 329–344 (1998)

[16] Li, L., Wang, Y.: Subjective trust inference in composite services. In: Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2010), pp. 1377–1384 (2010)

[17] Wang Y., Singh, M.P.: Evidence-based trust: A mathematical model geared for multiagent systems. ACM Transactions on Autonomous and Adaptive Systems 5(4), Article No. 14 (2010)

[18] Yao, Y.: Three-way decision: An interpretation of rules in rough set theory. In: Proceedings of the 4th International Conference on Rough Sets and Knowledge Technology (RSKT 2009), pp. 642–649 (2009)

[19] Chen, H., Tiňo, P., Yao, X.: Probabilistic classification vector machines. IEEE Transactions on Neural Networks 20(6), 901–914 (2009)

[20] Chen, H., Tiňo, P., Yao, X.: Predictive ensemble pruning by expectation propagation. IEEE Transactions Knowledge Data Engineering 21(7), 999–1013 (2009)

[21] Hamada, M.S., et al.: Bayesian reliability. Springer, Berlin (2008)

[22] Hines, W.W., et al.: Probability and statistics in engineering. John Wiley & Sons, Hoboken (2003)

[23] Chen, H., Tiňo, P., Yao, X.: Efficient probabilistic classification vector machine with incremental basis function selection. IEEE Transactions on Neural Networks and Learning Systems 25(2), 356–369 (2014)

[24] Jiang, B., et al.: Scalable graph-based semi-supervised learning through sparse bayesian model. IEEE Transactions on Knowledge and Data Engineering 29(12), 2758–2771 (2017)

[25] Cohen, W.W., Schapire, R.E., Singer, Y.: Learning to order things. Journal of Artificial Intelligence Research 5, 243–270 (1999)

[26] Goldberg, D., et al.: Using collaborative filtering to weave an information tapestry. Communications of the ACM 35(12), 61–70 (1992)

[27] Fieller, E., Herman, H., Pearson, E.: Tests for rank correlation coefficients. I. Biometrika, 44(3/4), 470–481 (1957)

[28] Gong, Z., Chen, H.: Model-based oversampling for imbalanced sequence classification. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM'16), pp. 1009–1018 (2016)

AUTHOR BIOGRAPHY

Lei Li received his Bachelor's degree from Jilin University, Changchun, China, in 2004, his Master's degree from the Memorial University of Newfoundland, St. John's, Canada, in 2006, and his Ph.D. degree from Macquarie University, Sydney, Australia, in 2012. He is currently an Associate Professor of computer science and technology at Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, China, Intelligent Interconnected Systems Laboratory of Anhui Province (Hefei University of Technology), and School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China. His research interests include data mining, social computing and graph computing. He is a senior member of IEEE.

ORCID: 0000-0002-5374-7293

Minghe Xue received the BE degree from Hefei University of Technology, China, in 2019. Her research interests are in graph computing and social computing. Currently, she is a Master candidate in Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, China, and School of Computer Science and Information Engineering at the Hefei University of Technology, Hefei, China.

ORCID: 0000-0001-8016-5136

Zan Zhang received his Ph.D. degree in Computer Science from Hefei University of Technology, China, in 2018. He is currently a lecturer at the Hefei University of Technology. His research interests include data mining and knowledge engineering.

ORCID: 0000-0002-6383-1683

Huanhuan Chen received the B.Sc. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2004, and the Ph.D. degree in computer science from the University of Birmingham, Birmingham, U.K., in 2008. He is currently a Full Professor with the School of Computer Science and Technology, USTC. His current research interests include neural networks, Bayesian inference, and evolutionary computation. He was the recipient of the 2015 International Neural Network Society Young Investigator Award, the 2012 IEEE Computational Intelligence Society Outstanding Ph.D. Dissertation Award, the IEEE Transactions on Neural Networks Outstanding Paper Award (bestowed in 2011 and only one paper in 2009), and the 2009 British Computer Society Distinguished Dissertations Award. He is currently an Associate Editor for the IEEE Transactions on Neural Networks and Learning System, and the IEEE Transactions on Emerging Topics in Computational Intelligence. He is a senior member of IEEE.
ORCID: 0000-0002-3918-384X

Xindong Wu is the foreign academician of Russian Academy of Science, and professor with the Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, China. He received his Bachelor's and Master's degrees in Computer Science from the Hefei University of Technology, China, and his Ph.D. degree in Artificial Intelligence from the University of Edinburgh, United Kingdom. His research interests include data mining, knowledge engineering, and Web information exploration. Dr. Wu is a Fellow of the IEEE and the AAAS. He is the Steering Committee Chair of the IEEE International Conference on Data Mining (ICDM), the Editor-in-Chief of Knowledge and Information Systems (KAIS, by Springer), and an Editor-in-Chief of the Springer Book Series on Advanced Information and Knowledge Processing (AI and KP). He was the Editor-in-Chief of the IEEE Transactions on Knowledge and Data Engineering (TKDE) between 2005 and 2008. He served as a program committee chair/co-chair for ICDM 2003 (the 3rd IEEE International Conference on Data Mining), KDD 2007 (the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining), CIKM 2010 (the 19th ACM Conference on Information and Knowledge Management, and ICBK 2017 (the 8th IEEE International Conference on Big Knowledge).
ORCID: 0000-0003-2396-1704